

Multiple-F0 estimation and music transcription

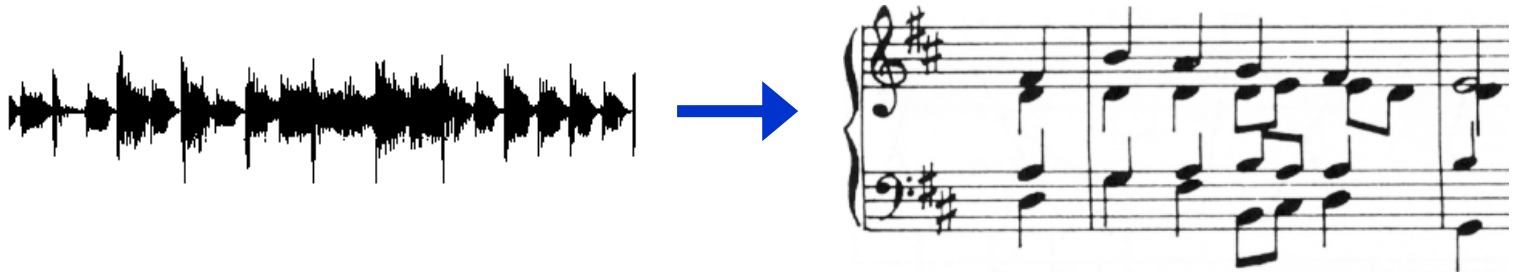
ISMIR Graduate School, October 4th-9th, 2004

Contents:

- Introduction
- Decomposition of the problem
- Multiple-F0 estimation
- Case study: TUT system

1 Introduction

- Music transcription
 - analyzing an acoustic musical signal so as to write down the parameters of the constituent sounds
- Using traditional musical notation:



- Automatic transcription in representational sense:
WAV → MIDI
- Written music is primarily a *performance instruction*, rather than a representation of music
 - discovering the "recipe", or, reverse-engineering the "source code" of a music signal

Previous work on music transcription

- Growing research interest during the last 10 years
- Larger scale projects:

University	People involved [when]
Stanford University	Moorer [75, 77], Chafe [82, 85]
University of Michigan	Piszczałski [79-86], Sterian [99]
Massachusetts Inst. of Tech.	Hawley [93], Martin [96]
University of Tokyo	Kashino [93, 95, 99]
AIST, Japan	Goto [95 – present]
University of London	Bello [03], Abdallah [03]
Cambridge University	Walmsley [99], Hainsworth [01], Davy [03]
Tampere University of Tech.	Klapuri, Eronen, Virtanen, Paulus, Ryyänen

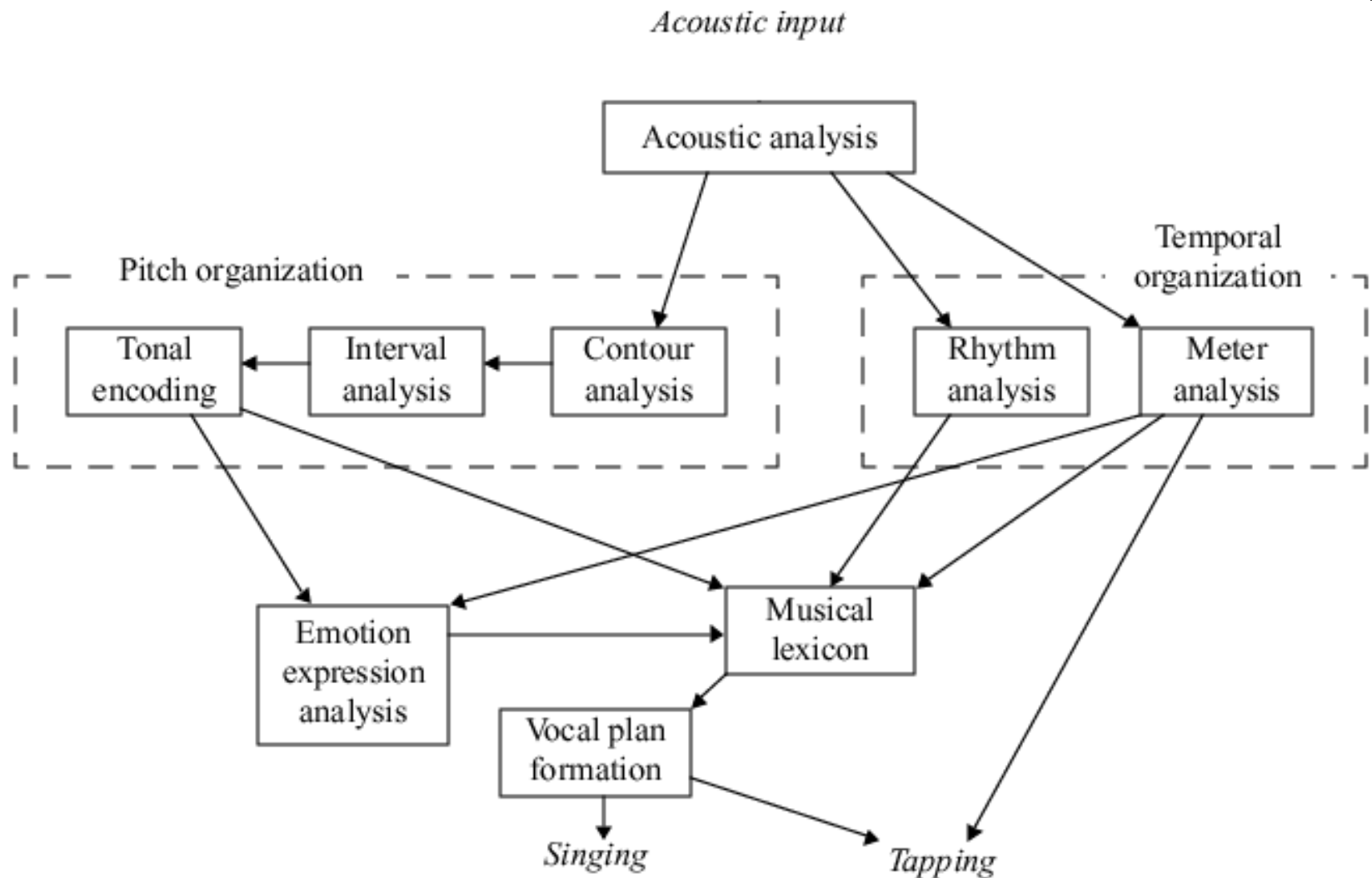
- Doctoral theses:
 - Moorer 75, Piszczałski 86, Maher 89, Mellinger 91, Hawley 93, Godsmark 98, Rossi 98, Sterian 99, Bello 03, Hainsworth 03, Klapuri04

State of the art

- General-purpose transcriber does not exist
 - accuracy and flexibility is not comparable to human musicians
 - even transcription of single-voice singing is not yet a solved problem: extracting *discrete notes* is difficult
- Some promising results for *limited-complexity music*
 - number of concurrent sounds limited
 - inference of drums and percussive instruments often not allowed
- Goto: extract melody and bass lines from real-world music

2 Decomposition of the transcription problem

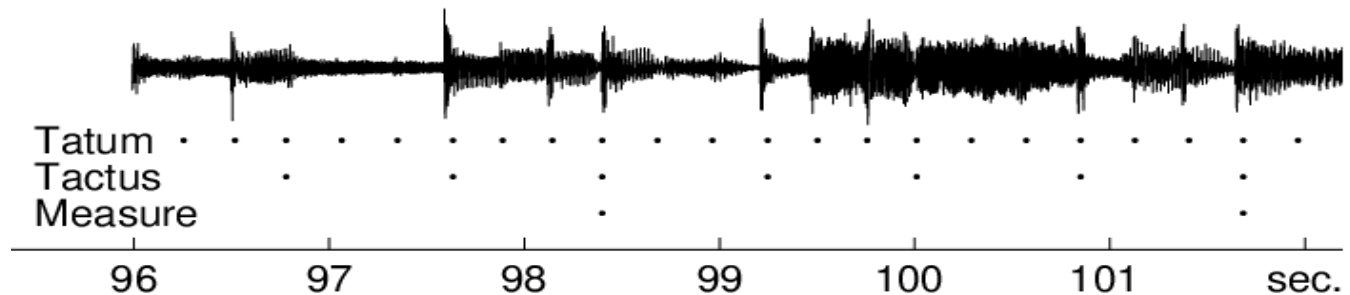
- Music transcription is a wide topic
- It is useful to structurize the problem by decomposing it into smaller and more tractable subproblems



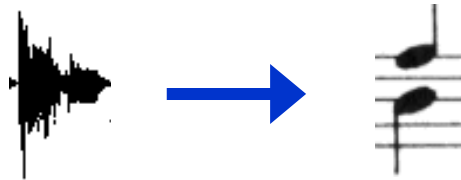
- *Functional modules of the music processing facility in the human brain as proposed by Peretz and Coltheart (Nature Neuroscience, 2003)*
 - *relative specialization to temporal / spectral resolution in the two hemispheres*

Meter analysis vs. pitch analysis

- Musical meter estimation
 - temporal segmentation at different time scales



- Multiple fundamental frequency (F0) estimation
 - find F0s of concurrent musical sounds
 - closely related to source separation



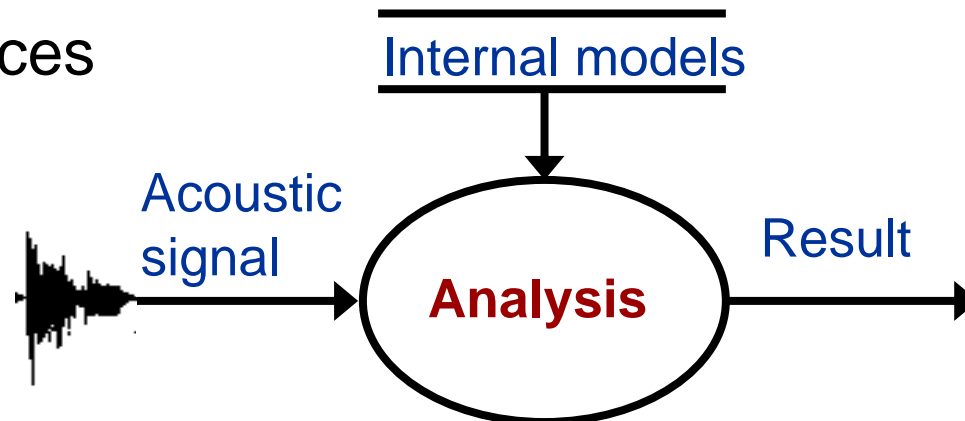
- Sound source recognition

Musicological models vs. acoustic signal

- Large-vocabulary speech recognition systems are critically dependent on *language models*
 - probabilities of different three-word sequences (N -gram models)
 - syntactic inference within sentences
- *Musicological information* is equally important for the AToM of polyphonically rich musical material
 - probabilities of different notes to occur simultaneously or concurrently
- Two main sources of information

$$P(\text{musical notation})$$

$$P(\text{musical notation})$$



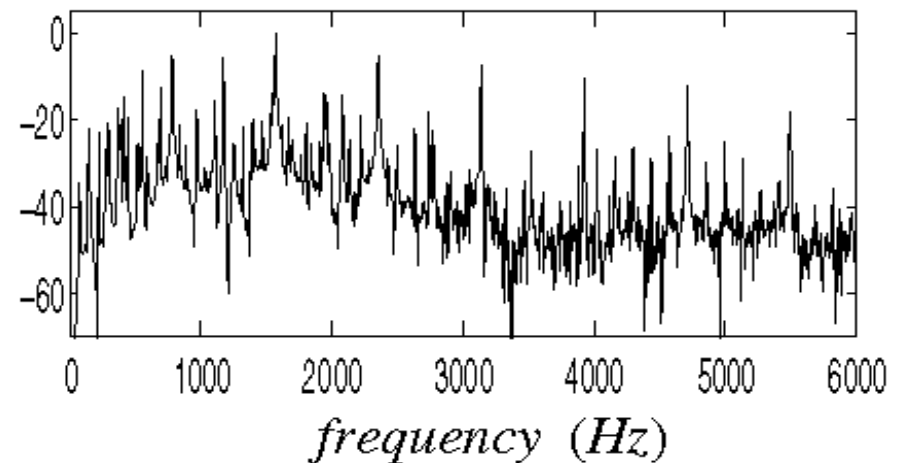
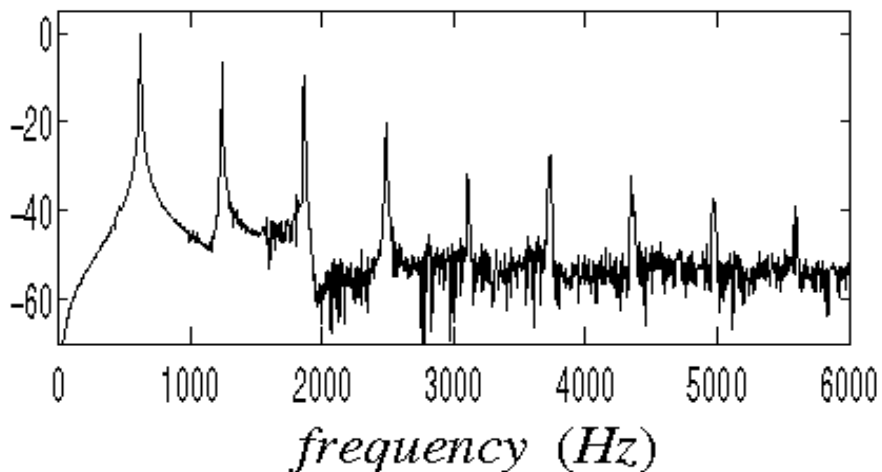
3 Multiple-F0 analysis

- Estimating the F0s of concurrent musical sounds
 - difficult problem
 - *music* → variety of sources; wide pitch range; presence of drums
- A number of completely different approaches have been proposed in the literature

Multiple-F0 vs. single-F0 estimation

- Complexity difference: see spectra below
- When two sounds overlap in time and frequency, there is no straightforward way of separating them again
→ different strategies introduced in the following

Spectrum of one vs. four concurrent harmonic sounds:



3.1 Perceptual grouping of time-frequency components

- An algorithm that finds the F0s of multiple concurrent sounds is also organizing spectral components to sources

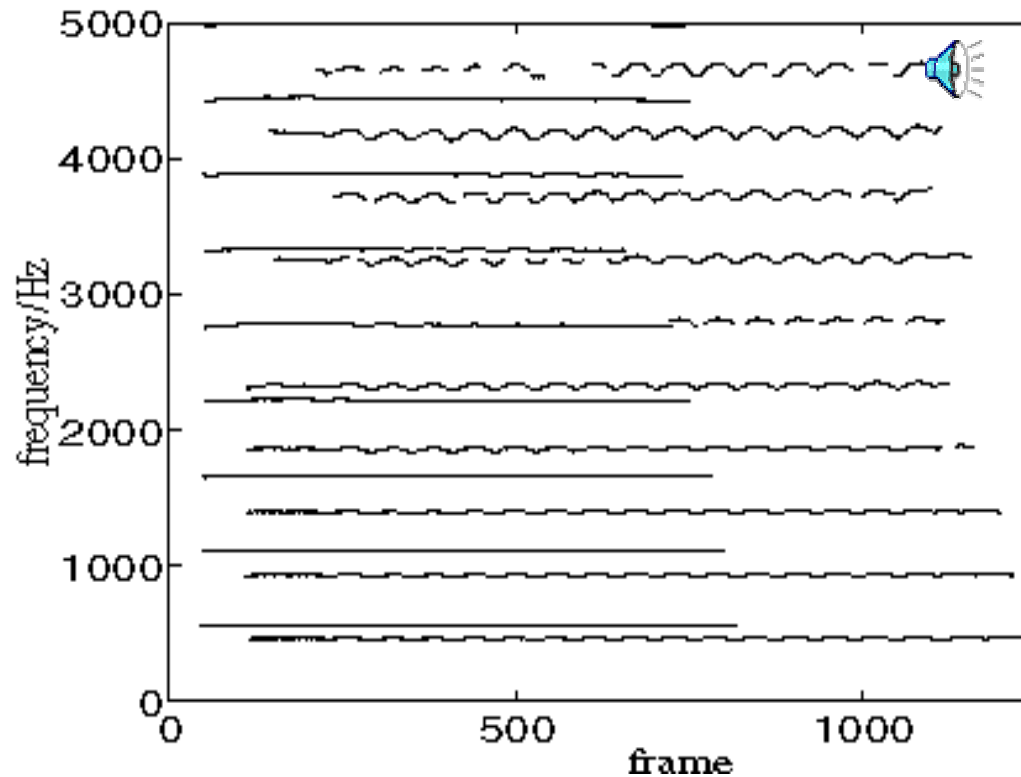
= *auditory scene analysis*

→ imitate human hearing

- *Perceptual cues* that promote component fusion [Bregman]

- proximity in time/freq
- harmonic relationships
- synchronous changes (onset, offset, AM, FM)
- spatial proximity

→ *Idea*: decompose, group



Perceptual grouping of time-frequency components

- Perceptual intentions of music oppose its transcription
- In music, different strategies are used to fuse component sounds into a coherent whole
 - *consonant musical intervals* are favoured to make the sounds "blend" better → a lot of frequency partials overlap
 - *several sounds set on simultaneously* (at metrically strong beats)
- In this sense, the transcription task is not "natural"
 - polyphonic transcription requires musical ear training

Perceptual grouping of time-frequency components

- Kashino *et al.* 1993, 1995
 - brought Bregman's ideas to music scene analysis
 - cluster sinusoidal partials using the *perceptual cues*
 - *higher-level internal models*
 - timbre models, tone memories
 - chord transition probabilities (trigram model)
 - *Bayesian probability network* for knowledge integration
 - bottom-up analysis, top-down predictions, temporal tying
 - evaluation material: 5 instruments, polyfonies around 3
- Sterian 1999 (PhD thesis)
 - Kalman filter to extract sinusoidal partials
 - perceptual grouping rules represented as likelihood functions
 - $P(\text{observed partials} \mid \text{note hypothesis})$
 - search strategy: multiple hypothesis tracking
 - evaluation material: few examples, polyphony 1-4 voices

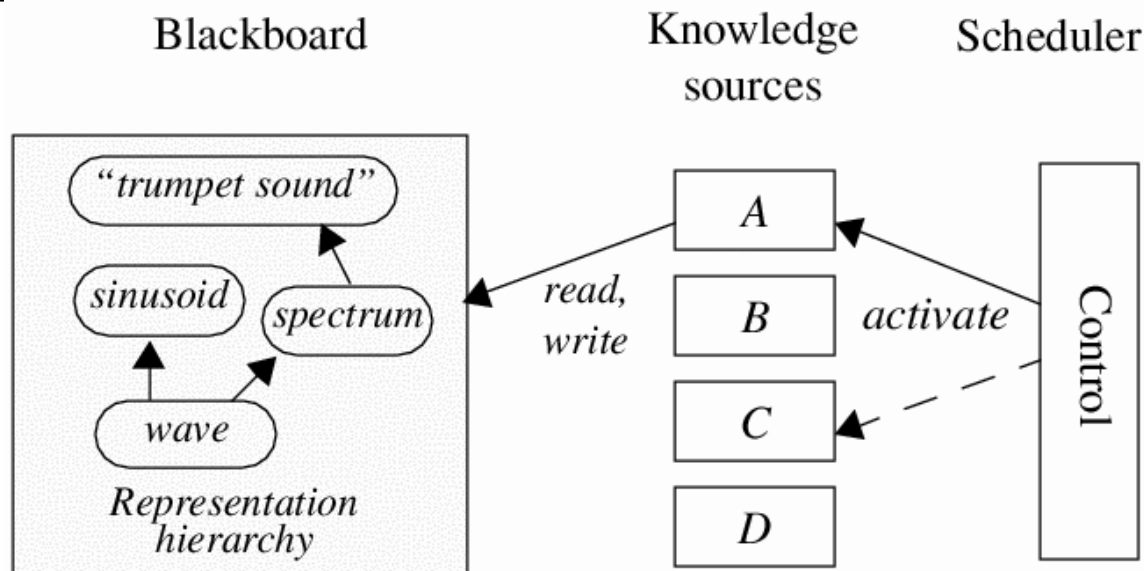
Auditory-model based approach

- Cheveigné and Kawahara 1993, 1999
 - extended the unitary pitch model to multiple-F0 case
 - *iterative approach*
 - pitch estimation is followed by the cancellation of the detected sound
 - evaluation results shown for rather artificial data
- Tolonen and Karjalainen 2000
 - *computationally efficient* implementation of the unitary pitch model
 - only two frequency channels instead of 40-120 in the original model
 - main characteristics of the model preserved
 - noise robustness considered
 - spectral flattening using inverse warped-LPC filtering
 - *non-iterative* extension to multipitch estimation
 - evaluation: method is rather accurate for F0s below 600 Hz

Auditory-model + musicological rules

■ Martin 1996

- system for transcribing four-voice Bach chorales
- *auditory model* was used as a front-end
- blackboard architecture was used to integrate knowledge of physical sound production with *musical rules* governing tonal music



3.3 Signal-model based probabilistic inference

- The whole multiple-F0 estimation problem can be stated in terms of a *signal model, the parameters of which should be estimated*
- Consider the model

$$y_t = \left\{ \sum_{k=1}^K \sum_{m=1}^{M_k} a_{k,m} \cos[m\omega_k t] + b_{k,m} \sin[m\omega_k t] \right\} + v_t$$

K : number of simultaneous sounds

M_k : number of partials in note k

ω_k : fundamental frequency of note k

$a_{k,m}, b_{k,m}$: represent the amplitude and phase of a partial

v_t : residual noise

- In principle, all the right-hand side parameters should be estimated based on the observation y_t and possible prior knowledge on the parameter distributions → *Bayes*

Signal-model based probabilistic inference

$$y_t = \left\{ \sum_{k=1}^K \sum_{m=1}^{M_k} a_{k,m} \cos[m\omega_k t] + b_{k,m} \sin[m\omega_k t] \right\} + v_t$$

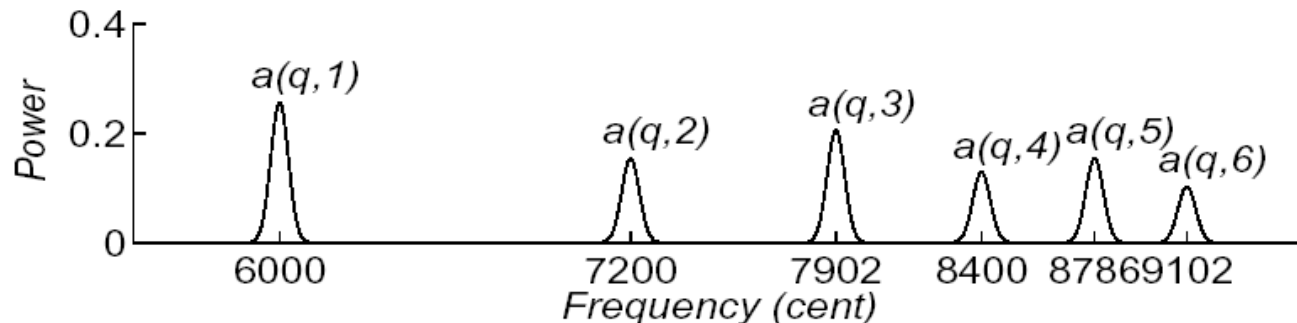
■ Davy and Godsill 2003

- augmented the above signal model
 - time-varying amplitudes; non-ideal harmonicity; non-white residual noise
- prior distributions for the parameters
- parameter estimation *in time domain*
- *problems*: parameter space is huge, posterior distribution is highly multimodal
- Markov chain Monte Carlo (MCMC) sampling of the posterior
- evaluation: robust for polyphonies of up to 3 sounds, slow to compute

Signal-model based probabilistic inference

■ Goto

- *short-time spectrum* of a signal is modeled as a weighted mixture of tone models
- *EM-algorithm*
 - iterative updating of tone models and their weights
 - maximum *a posteriori* (MAP) estimate
- temporal continuity was considered using multiple tracking agents
- used to detect the melody and bass lines in CD recordings
- EM-algorithm can be easily implemented based on reference



3.4 Data-adaptive techniques

- *No* parametric model or other knowledge of the sources
 - source signals are estimated from the data
 - typically, sources are not even assumed to have harmonic spectra!
- For real-world signals, e.g. ICA performs poorly
- Restrictions for the sources
 - data-adaptive techniques become applicable in realistic cases
 - *independence* of the sources
 - *sparseness*, meaning that sources are inactive most of the time

Data-adaptive techniques

■ Virtanen 2003

– constraints for the sources:

- *sparseness*
- *temporal continuity*

– signal model

$$X_t(f) = \sum_{n=1}^N a_{t,n} S_n(f) + E_t(f)$$

$X_t(f)$: power spectrogram of the input

$S_n(f)$: power spectrum of source n

$a_{t,n}$: time-varying gains of the sources

$E_t(f)$: error spectrogram

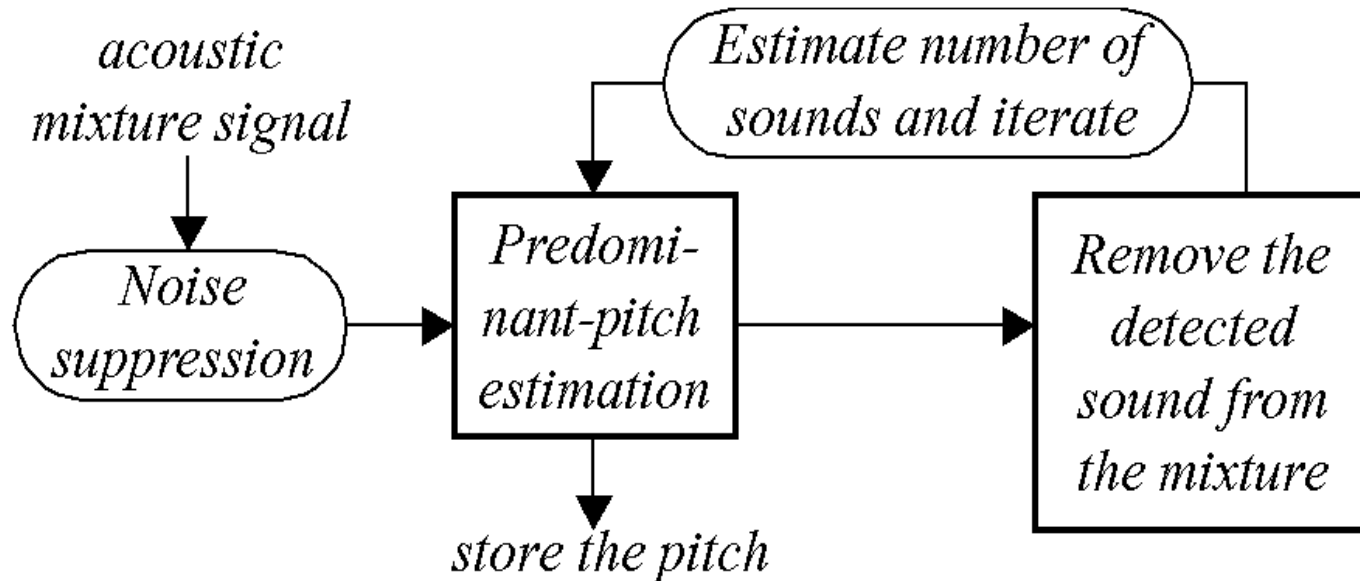
- *iterative optimization algorithm* which finds non-negative $a_{t,n}$ and $S_n(f)$
- evaluation: separation of pitched and drum instruments from real-world signals

■ Abdallah and Plumbley [unpublished]

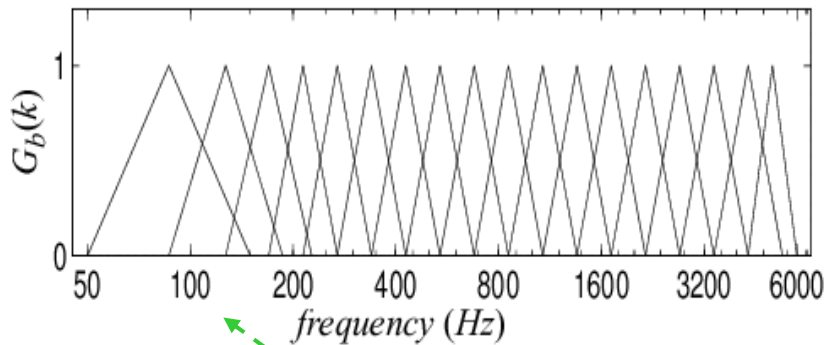
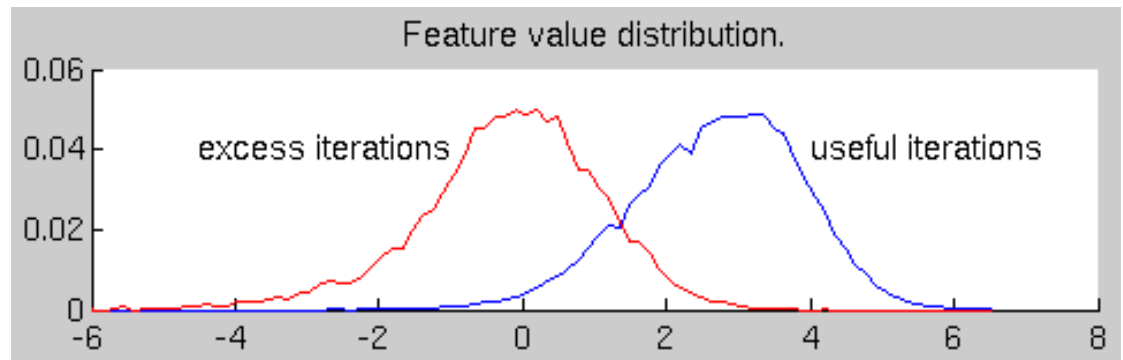
- in many ways similar to above

4 Case study: TUT method

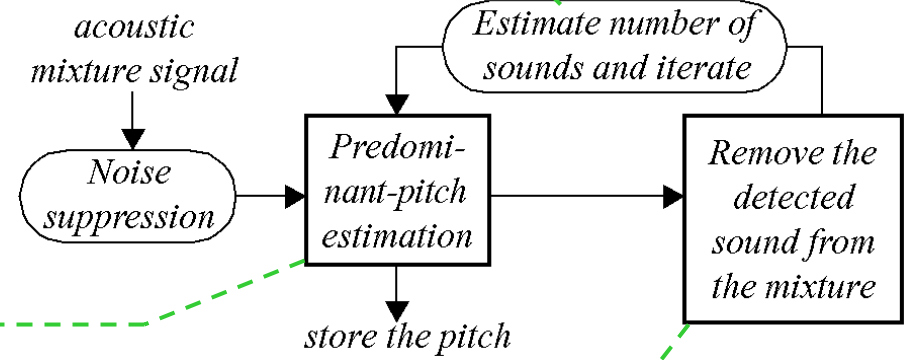
- Multiple-F0 estimation strategy [Klapuri, 2003]
 1. estimate one sound in a mixture signal
 2. cancel the detected sound
 3. repeat estimation for the residual



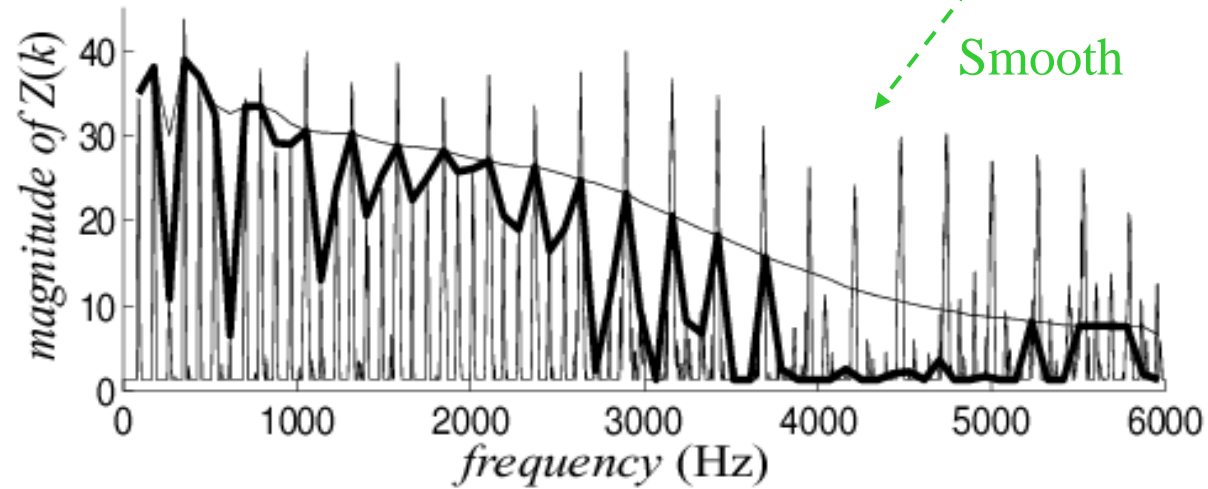
Case study: TUT method



Bandwise processing



Feature statistics







Smooth



Transcription demonstrations

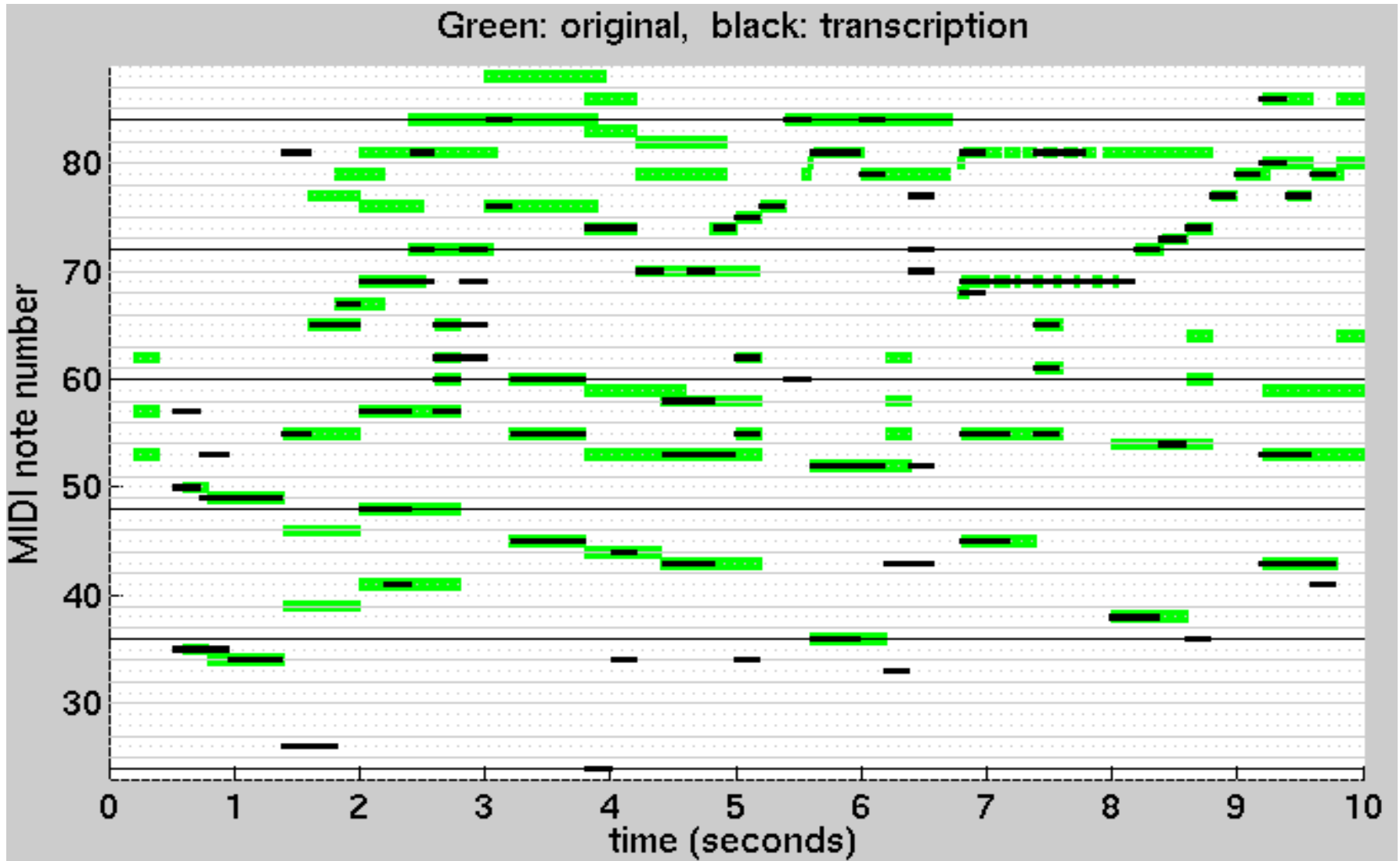
(Available at <http://www.cs.tut.fi/~klap/iiro/>)

Original	Transcribed, resynthesized	Comments
Chopin: Nocturne 		piano music
Larrissey: Mist on the mountain 		no drums
Zuccherro: Senza una donna 		easy pop music
Abba: Dancing queen 		pop music
Taylor: Lady my whole world is 		pop music
Saint-Saens: Sarabande op. 93 		classical
Beethoven: 5th symphony 		classical
Kool and the Gang: Funky stuff 		funk: difficult
The Police: It's alright for you 		rock: difficult

Transcription of synthesized MIDI

- *Some evaluation of a transcription system* can be done by transcribing synthesized MIDI-songs
 - exact reference score is available
 - high-quality MIDI-songs can be thought to simulate the real case
- Pure listening of transcription results is not sufficiently informative
 - some errors are not heard in listening (e.g. note omissions, octave errors)
 - few *aurally* bad errors may overshadow otherwise good job
- Synthesized MIDI in general: slightly easier than signals on CDs

Piano examples

*Georgia*

5 Conclusions

- *Multipitch estimation* is possible to some degree, even without source models and in a single time frame
- In continuous music, however
 - drum sounds are present
 - polyphony is high and not known *a priori*
- It seems that *musicological (higher-level) models* are necessary to further improve the transcription accuracy
 - system must understand something about music, not just listen